Материалы к лекции «Уменьшение размерности в данных. Метод главных компонент» по курсу «Математические основы теории прогнозирования» 2010

Рассмотрим задачу классификации изображений рукописных цифр MNIST[1]. Пусть имеется некоторое количество черно-белых изображений, на каждом из которых представлена одна цифра (см. рис. 1). Задача состоит в автоматическом определении цифры для входного изображения.

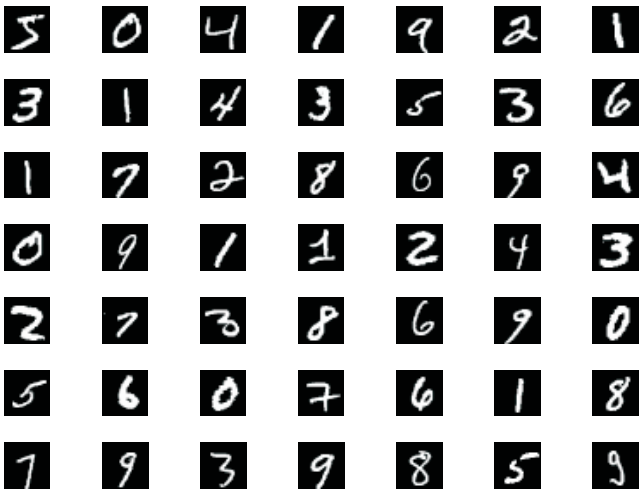


Рис. 1: Примеры изображений рукописных цифр из базы данных MNIST.

Для того, чтобы применить методы распознавания в данной задаче, необходимо предварительно выбрать пространство признаков, характеризующее изображения цифр. В простейшей случае в качестве признаков можно взять исходные интенсивности пикселов изображения. Тогда для изображения размера 28×28 получаем 784 признака. Такой способ формирования признакового пространства обладает рядом существенных недостатков. Во-первых, получается большое количество признаков. Например, для относительно небольших изображений размера 300×200 получается 60000 признаков. Большое количество признаков приводит к высоким временным затратам на обработку данных, большим объемам памяти, требуемой для хранения информации, а также к необходимости сбора большого числа прецедентов для уверенного восстановления скрытых зависимостей в существенно многомерном

[1]Исходные данные можно скачать по адресу http://yann.lecun.com/exdb/mnist/

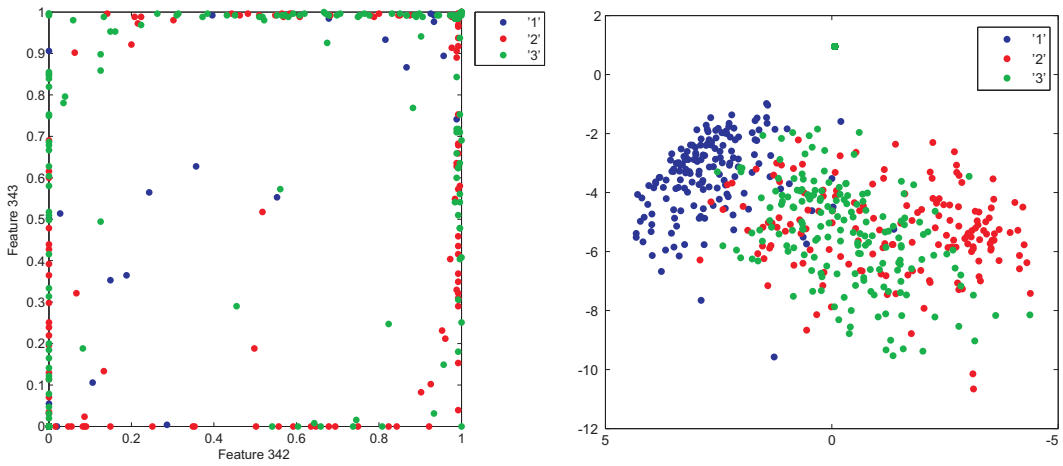


Рис. 2: Проекция выборки изображений цифр ’1’, ’2’ и ’3’ на два признака, соответствующих интенсивностям пикселов (a) и на два признака, полученных с помощью метода главных компонент (b).

пространстве. Другим серьезным недостатком полученного признакового пространства является тот факт, что близкие в пространстве признаков объекты не соответствуют одним и тем же классам (см. рис. 2a). Выполнение гипотезы компактности является одним из основных требований для большинства методов распознавания. Методы уменьшения размерности в данных позволяют получать представление выборок в маломерных пространствах, обладающих рядом хороших свойств. В частности, для изображений рукописных цифр метод главных компонент позволяет получить существенно более качественное признаковое пространство (см. рис. 2b).

Пусть имеется некоторая выборка объектов $X = \{\boldsymbol{x}_n\}_{n=1}^N$, $\boldsymbol{x}_n \in \mathbb{R}^D$. Задача уменьшения размерности состоит в получении представления этой выборки в пространстве меньшей размерности $T = \{\boldsymbol{t}_n\}_{n=1}^N, \boldsymbol{t}_n \in \mathbb{R}^d$. Здесь $d \ll D$, но в частных случаях d может и совпадать с D. Уменьшение размерности в описании данных может преследовать множество целей:

- Сокращение вычислительных затрат при обработке данных;
- Борьба с переобучением. Чем меньше количество признаков, тем меньше требуется объектов для уверенного восстановления скрытых зависимостей в данных и тем больше качество восстановления подобных зависимостей;
- Сжатие данных для более эффективного хранения информации. В этом случае помимо преобразования $X \to T$ требуется иметь возможность осуществлять также обратное преобразование $T \to X$;
- Визуализация данных. Проектирование выборки на двух-/трехмерное пространство позволяет графически представить выборку;
- Извлечение новых признаков. Новые признаки, полученные в результате преобразования $X \to T$, могут оказывать значимый вклад при последующем решении задач распознавания (например, как метод главных компонент в случае рис. 2b);

- и др.

Заметим, что все описанные далее методы уменьшения размерности относятся к классу методов обучения без учителя, т.е. в качестве исходной информации выступает только признаковое описание объектов X. В частности, в задаче классификации рукописных цифр результат, показанный на рис. 2b, был получен без использования информации о цифрах.

1 Метод главных компонент

Метод главных компонент (разложение Карунена-Лоева, principal component analysis, PCA) является простейшим методом уменьшения размерности в данных, идеи которого высказывались еще в 19 веке [1, 2]. Идея метода заключается в поиске в исходном пространстве гиперплоскости заданной размерности с последующим проектированием выборки на данную гиперплоскость. При этом выбирается та гиперплоскость, ошибка проектирования данных на которую является минимальной в смысле суммы квадратов отклонений.

Рассмотрим произвольный ортонормированный базис в пространстве $\mathbb{R}^D$: $\boldsymbol{w}_1, \ldots, \boldsymbol{w}_D$, $\boldsymbol{w}_i^T \boldsymbol{w}_j = \delta_{ij}$, где $\delta_{ij} = [i = j]$ – символ Кронекера. Не ограничивая общности, будем считать, что первые d векторов этого базиса $\boldsymbol{w}_1, \ldots, \boldsymbol{w}_d$ образуют базис искомой гиперплоскости. Тогда точки гиперплоскости определяются как

$$\boldsymbol{x} = \boldsymbol{w}_1 t_1 + \cdots + \boldsymbol{w}_d t_d + \boldsymbol{\mu} = W\boldsymbol{t} + \boldsymbol{\mu},$$

где $t_1, \ldots, t_d$ – координаты точки $\boldsymbol{x}$ в базисе гиперплоскости, $W = (\boldsymbol{w}_1 | \ldots | \boldsymbol{w}_d) \in \mathbb{R}^{D \times d}$ – матрица, столбцы которой представляют собой базисные вектора $\boldsymbol{w}_1, \ldots, \boldsymbol{w}_d$, $\boldsymbol{\mu} \in \mathbb{R}^D$ – вектор сдвига. Поиск гиперплоскости, обеспечивающей минимальную квадратичную ошибку проектирования, может быть записан как

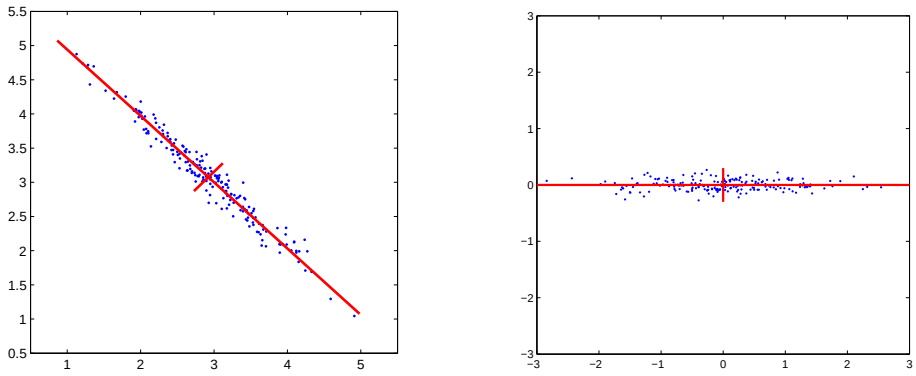
$$J = \sum_{n=1}^{N} \|\boldsymbol{x}_n - W\boldsymbol{t}_n - \boldsymbol{\mu}\|^2 \to \min_{\boldsymbol{w}_1, \ldots, \boldsymbol{w}_d, \boldsymbol{t}_1, \ldots, \boldsymbol{t}_d, \boldsymbol{\mu}}. \tag{1}$$

Введем следующие обозначения:

$$\bar{\boldsymbol{x}} = \frac{1}{N} \sum_{n=1}^{N} \boldsymbol{x}_n \text{ – выборочное среднее,}$$

$$S = \frac{1}{N} \sum_{n=1}^{N} (\boldsymbol{x}_n - \bar{\boldsymbol{x}})(\boldsymbol{x}_n - \bar{\boldsymbol{x}})^T \text{ – выборочная матрица ковариации.}$$

Рассмотрим собственные вектора и собственные значения матрицы S: $S = Q\Lambda Q^T$, где $\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_D)$, Q – ортогональная матрица ($Q^T Q = I$), в столбцах которой стоят собственные вектора. Предположим, без ограничения общности, что собственные значения отсортированы по убыванию, т.е. $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_D$. Можно показать [1, 2] (см. Приложение A), что задача оптимизации (1) имеет следующее аналитическое решение: $\boldsymbol{w}_1, \ldots, \boldsymbol{w}_d$ – собственные вектора матрицы S, отвечающие d наибольшим собственным значениям $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_d$, $\boldsymbol{\mu} = \bar{\boldsymbol{x}}$, $\boldsymbol{t}_n = W^T(\boldsymbol{x}_n - \bar{\boldsymbol{x}})$. При этом значение критерия J в точке минимума равно $\sum_{i=d+1}^{N} \lambda_i$. Таким образом, величина ошибки проектирования для оптимальной гиперплоскости составляет сумму дисперсий данных по отбрасываемым размерностям, определяемым собственными векторами $\boldsymbol{w}_{d+1}, \ldots, \boldsymbol{w}_D$.



(a) (b)

Рис. 3: Пример применения метода главных компонент. На рис. a показана исходная выборка в двухмерном пространстве вместе с направлениеми, определяемыми собственными векторами выборочной матрицы ковариации. На рис. b показан переход к некоррелированным признакам.

Наряду с критерием минимизации ошибки проектирования можно рассмотреть альтернативный критерий поиска гиперплоскости, связанный с максимизацией разброса спроектированных точек выборки. Пусть $\hat{S}$ – выборочная матрица ковариации для спроектированных точек выборки. Тогда величину разброса можно определить как $\mathrm{tr}(\hat{S})$. В одномерном случае этот критерий совпадает с дисперсией данных. Можно показать, что решение задачи

$$\mathrm{tr}(\hat{S}) \to \max_{\boldsymbol{w}_1,\ldots,\boldsymbol{w}_d,\boldsymbol{t}_1,\ldots,\boldsymbol{t}_d,\boldsymbol{\mu}}$$

совпадает с решением задачи (1). При этом для оптимальной гиперплоскости величина критерия $\mathrm{tr}(\hat{S}) = \sum_{i=1}^{d} \lambda_i$, где, как и раньше, λ_i – собственные значения выборочной матрицы ковариации S для исходной выборки.

Итак, метод главных компонент предполагает переход от исходного базиса к базису из собственных векторов матрицы ковариации S с дальнейшим отбрасыванием проекций выборки на собственные вектора, отвечающие $D - d$ наименьшим собственным значениям. В базисе из собственных векторов матрица ковариации S имеет диагональный вид $\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_D)$. Таким образом, признаки, получаемые с помощью метода главных компонент, являются некоррелированными. Переход к некоррелированным признакам часто является разумным методом предобработки исходных данных. Поэтому метод главных компонент применяется и в случае $d = D$.

Рассмотрим простой модельный пример применения метода главных компонент. Пусть исходная выборка представляет собой данные в двухмерном пространстве (см. рис. 3a). При использовании метода главных компонент центр координат нового пространства переносится в центр выборки, а оси определяются собственными векторами выборочной матрицы ковариации (см. рис. 3b). Таким образом, новые признаки являются некоррелированными. В том случае, если $d = 1$, то дополнительно осуществляется проекция выборки на направление, соответствующее наибольшему собственному значению (направление с наибольшей дисперсией). Для данного примера это координата x.

Гиперплоскость, которая находится с помощью метода главных компонент, определена однозначно в том случае, если $\lambda_d > \lambda_{d+1}$. При этом остается произвол в выборе системы

Алгоритм 1 Схема метода главных компонент

Вход: $X \in \mathbb{R}^{N \times D}$ – исходная выборка данных, d – размерность редуцированного пространства
Выход: $T \in \mathbb{R}^{N \times d}$ – представление выборки в редуцированном пространстве

$\bar{\boldsymbol{x}} = \frac{1}{N} \sum_{n=1}^{N} \boldsymbol{x}_n$; *// Вычисляем выборочное среднее*
$\boldsymbol{x}_n \leftarrow \boldsymbol{x}_n - \bar{\boldsymbol{x}}$; *// Переносим начало координат в центр выборки*
если $N > D$ **то**
  $S = \frac{1}{N} X^T X$; *// Вычисляем выборочную матрицу ковариации*
  $S = Q \Lambda Q^T, \Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_D), Q^T Q = I, Q = (\boldsymbol{q}_1 | \ldots | \boldsymbol{q}_D)$; *// Находим собственные вектора и собственные значения матрицы ковариации*
  Выбираем d наибольших собственных значений $\lambda_1 \geq \lambda_2 \geq \ldots \lambda_d$ и соответствующие им собственные вектора $W = (\boldsymbol{q}_1 | \ldots | \boldsymbol{q}_d)$;
иначе
  $S = \frac{1}{N} X X^T$;
  $S = Q \Lambda Q^T, \Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_D), Q^T Q = I, Q = (\boldsymbol{q}_1 | \ldots | \boldsymbol{q}_D)$; *// Находим собственные вектора и собственные значения матрицы S*
  $Q \leftarrow \frac{1}{\sqrt{N}} X^T Q \mathrm{diag}\left(\frac{1}{\sqrt{\lambda_1}}, \ldots, \frac{1}{\sqrt{\lambda_D}}\right)$; *// Переходим к нормированным собственным векторам выборочной матрицы ковариации*
  Выбираем собственные вектора, соответствующие d наибольшим собственным значениям $W = (\boldsymbol{q}_1 | \ldots | \boldsymbol{q}_d)$;
$T = XW$; *// Проектируем выборку на выбранные направления*

координат в линейном пространстве, определяемом гиперплоскостью, т.е. выборе значений T. Этот выбор не оказывает влияние на величину ошибки при восстановлении данных, которая определяется только самой гиперплоскостью. Стандартный способ выбора базиса в пространстве гиперплоскости – собственные вектора матрицы ковариации с нулем в центре выборки. Однако, в ряде случаев такой выбор базиса является неадекватным. Например, метод независимых компонент (см. [6]) представляет собой другой способ выбора базиса гиперплоскости, который активно применяется для задачи разделения независимых источников.

При использовании метода главных компонент необходимо вычислять собственные вектора и собственные значения выборочной матрицы ковариации, которая имеет размер $D \times D$. Сложность этой операции составляет $O(D^3)$. В том случае, если $D > N$, то существует способ более экономного вычисления собственных векторов и собственных значений матрицы ковариации с помощью матрицы размера $N \times N$ и сложностью, соответственно, $O(N^3)$. Действительно, в пространстве размерности D множество из N точек образует линейное подпространство максимальной размерности $N-1$. Поэтому не имеет смысла применять метод главных компонент для $d > N - 1$. С точки зрения матрицы ковариации это означает, что только $N - 1$ собственных значений отличны от нуля. Все остальные собственные вектора не имеет смысла вычислять, т.к. дисперсия выборки вдоль этих направлений равна нулю.

Пусть $X \in \mathbb{R}^{N \times D}$ – исходная выборка с нулевым центром, т.е. $\frac{1}{N} \sum_{n=1}^{N} \boldsymbol{x}_n = \boldsymbol{0}$. Тогда выборочная матрица ковариации $S = \frac{1}{N} X^T X$. Рассмотрим собственные вектора и собственные значения матрицы S:

$$\frac{1}{N} X^T X \boldsymbol{q}_i = \lambda_i \boldsymbol{q}_i.$$

Домножим обе части этого уравнения на X:

$$\frac{1}{N}XX^T(X\boldsymbol{q}_i) = \lambda_i(X\boldsymbol{q}_i).$$

Обозначая $\boldsymbol{v}_i = X\boldsymbol{q}_i$, получаем

$$\frac{1}{N}XX^T\boldsymbol{v}_i = \lambda_i\boldsymbol{v}_i.$$

Таким образом, матрица $\frac{1}{N}XX^T$ размера $N \times N$ имеет те же собственные значения, что и выборочная матрица ковариации S (у которой, в свою очередь, есть $D - N$ дополнительных нулевых собственных значений). Сложность поиска собственных значений и собственных векторов матрицы $\frac{1}{N}XX^T$ составляет $O(N^3)$, что может давать значительную выгоду по сравнению с $O(D^3)$ при $D > N$. Для получения собственных векторов матрицы S домножим обе части последнего уравнения на X^T:

$$\frac{1}{N}X^TX(X^T\boldsymbol{v}_i) = \lambda_i(X^T\boldsymbol{v}_i).$$

Таким образом, $X^T\boldsymbol{v}_i$ является собственным вектором матрицы S, отвечающим собственному значению λ_i. Однако, в том случае, если исходные вектора $\boldsymbol{v}_i$ являются нормированными, т.е. $\|\boldsymbol{v}_i\| = 1$, то вектора $X^T\boldsymbol{v}_i$ нормированными уже не являются. Нормированные вектора можно получить с помощью следующего выражения:

$$\boldsymbol{q}_i = \frac{1}{\sqrt{N\lambda_i}}X^T\boldsymbol{v}_i.$$

Теперь, объединяя все вышесказанное, можно составить схему метода главных компонент, представленную в алгоритме 1.

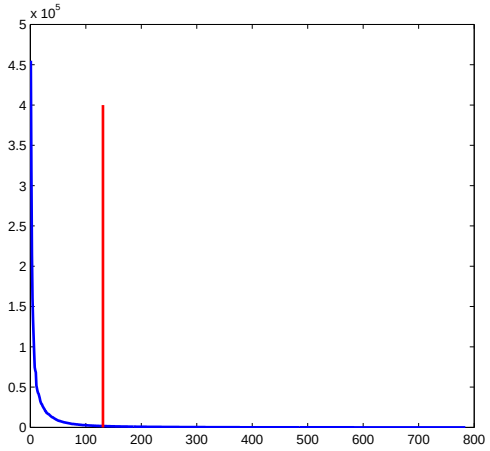


Рис. 4: Схема выбора размерности редуцированного пространства для метода главных компонент.

До сих пор предполагалось, что размерность редуцированного пространства d задается пользователем заранее. Это значение легко выбрать в том случае, если стоит задача визуализации данных ($d = 2$ или $d = 3$) или задача вложения выборки в заданный объем

памяти. Однако, во многих других случаях выбор d является далеко не очевидным из априорных предположений. Для метода главных компонент существует простой эвристический прием выбора величины d. Одной из особенностью метода главных компонент является тот факт, что все редуцированные пространства для $d = 1, 2, \ldots$ являются вложенными друг в друга. В частности, однократное вычисление всех собственных векторов и собственных значений матрицы ковариации позволяет получить редуцированное пространство для любого значения d. При этом ошибка проектирования данных на соответствующую гиперплоскость составляет $\sum_{i=d+1}^{D} \lambda_i$. Поэтому для выбора значения d можно отобразить на графике собственные значения в порядке убывания (см. рис. 4) и выбрать порог отсечения таким образом, чтобы справа остались значения, не значимо отличные от нуля. Другой способ предполагает выбор порога так, чтобы справа оставался определенный процент от общей площади под кривой (например, 5% или 1%), т.е.

$$d : \frac{\sum_{i=d+1}^{D} \lambda_i}{\sum_{i=1}^{D} \lambda_i} < \eta.$$

Площадь под кривой определяется значением $\mathrm{tr}(S)$ и соответствует величине разброса в данных.

2 Вероятностный метод главных компонент

Для метода главных компонент можно сформулировать вероятностную модель (probabilistic PCA, PPCA) (см. [3]). Переформулирование метода в вероятностных терминах дает целый ряд преимуществ, а именно:

- Возможность использования EM-алгоритма для поиска решения. Для PCA EM-алгоритм является вычислительно более эффективной процедурой в ситуациях, когда $d \ll D$;
- Корректная обработка пропущенных значений. Пропущенные значения просто добавляются в список скрытых переменных вероятностной модели, для которой затем применяется соответствующий EM-алгоритм;
- Возможность перехода к модели смеси вероятностных распределений, которая значительно расширяет область применимости метода (см. [7]);
- Возможность использования т.н. байесовского подхода для решения задач выбора моделей и, в частности, здесь можно построить теоретически обоснованную схему выбора размерности редуцированного пространства d (см. [4, 5]);
- Возможность генерирования новых объектов из вероятностной модели;
- Для задач классификации возможность моделирования распределений отдельных классов объектов для дальнейшего использования в различных схемах классификации;
- Значение функции правдоподобия является универсальным критерием, позволяющим сравнивать различные вероятностные модели между собой. В частности, в помощью значения правдоподобия можно легко определять выбросы в данных.

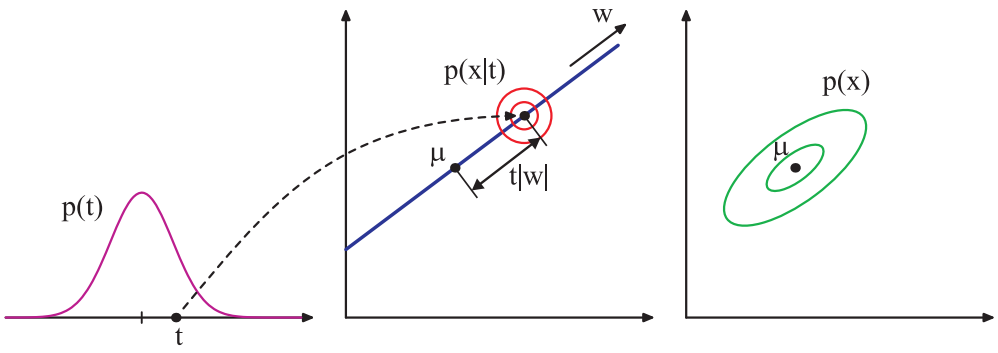


Рис. 5: Иллюстрация процесса генерации объекта в вероятностной модели PCA для $D = 2$ и $d = 1$. Наблюдаемое значение $\boldsymbol{x}$ образуется путем генерирования значения скрытой компоненты t из априорного распределения $p(t)$ и последующего генерирования значения $\boldsymbol{x}$ из изотропного нормального распределения с центром $\boldsymbol{\mu} + t\boldsymbol{w}$ и матрицей ковариации $\sigma^2 I$. Зеленые эллипсы показывают линии уровня плотности маргинального распределения $p(\boldsymbol{x})$.

Сформулируем вероятностную модель PCA следующим образом:

$$\boldsymbol{x} = W\boldsymbol{t} + \boldsymbol{\mu} + \boldsymbol{\varepsilon}.$$

Здесь, как и раньше, $W\boldsymbol{t} + \boldsymbol{\mu}$ задает точку на гиперплоскости, а $\boldsymbol{\varepsilon} \sim \mathcal{N}(\boldsymbol{\varepsilon}|\mathbf{0}, \sigma^2 I)$ – нормально распределенная шумовая компонента с одинаковой дисперсией σ^2 по всем направлениям в пространстве $\mathbb{R}^D$. Символом $\mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}, \Sigma)$ здесь и далее будет обозначаться плотность многомерного нормального распределения:

$$\mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}, \Sigma) = \frac{1}{\sqrt{2\pi}^D \sqrt{\det \Sigma}} \exp\left(-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu})^T \Sigma^{-1} (\boldsymbol{x} - \boldsymbol{\mu})\right).$$

В качестве априорного распределения на значение координат объекта $\boldsymbol{t}$ в базисе гиперплоскости выберем следующее:

$$p(\boldsymbol{t}) = \mathcal{N}(\boldsymbol{t}|\mathbf{0}, I).$$

Процесс генерации объекта $\boldsymbol{x}$ в заданной вероятностной модели показан на рис. 5. Полное совместное распределение в вероятностной модели PCA задается следующим образом:

$$p(X, T|W, \boldsymbol{\mu}, \sigma) = \prod_{n=1}^{N} p(\boldsymbol{x}_n, \boldsymbol{t}_n|W, \boldsymbol{\mu}, \sigma) = \prod_{n=1}^{N} \mathcal{N}(\boldsymbol{x}_n|W\boldsymbol{t}_n + \boldsymbol{\mu}, \sigma^2 I)\mathcal{N}(\boldsymbol{t}_n|\mathbf{0}, I).$$

Здесь X – набор наблюдаемых переменных, T – набор скрытых переменных и $(W, \boldsymbol{\mu}, \sigma)$ – набор параметров модели.

Для поиска значений параметров модели воспользуемся методом максимального правдоподобия:

$$p(X|W, \boldsymbol{\mu}, \sigma) = \prod_{n=1}^{N} p(\boldsymbol{x}_n|W, \boldsymbol{\mu}, \sigma) \to \max_{W, \boldsymbol{\mu}, \sigma}. \tag{2}$$

Маргинальное распределение $p(\boldsymbol{x}_n)$ в вероятностной модели PCA вычисляется как

$$p(\boldsymbol{x}_n|W, \boldsymbol{\mu}, \sigma) = \int p(\boldsymbol{x}_n|\boldsymbol{t}_n, W, \boldsymbol{\mu}, \sigma) p(\boldsymbol{t}_n) d\boldsymbol{t}_n.$$

Последний интеграл представляет собой свертку двух нормальных распределений и может быть вычислен аналитически:

$$p(\boldsymbol{x}_n|W,\boldsymbol{\mu},\sigma) = \int \mathcal{N}(\boldsymbol{x}_n|W\boldsymbol{t}_n+\boldsymbol{\mu},\sigma^2 I)\mathcal{N}(\boldsymbol{t}_n|\boldsymbol{0},I)d\boldsymbol{t}_n = \mathcal{N}(\boldsymbol{x}_n|\boldsymbol{\mu},\sigma^2 I + WW^T).$$

Таким образом, вероятностная модель PCA представляет собой нормальное распределение, в котором матрица ковариаций задается специальным образом

$$C = WW^T + \sigma^2 I. \tag{3}$$

Заметим, что также как и классическая модель PCA, вероятностная модель PCA инвариантна относительно выбора базиса в гиперплоскости. Пусть $R \in \mathbb{R}^d$ – произвольная ортогональная матрица, задающая поворот базиса гиперплоскости. Это соответствует использованию матрицы $\widetilde{W} = WR$. Тогда матрица ковариаций равна

$$C = \widetilde{W}\widetilde{W}^T + \sigma^2 I = WRR^TW^T + \sigma^2 I = WW^T + \sigma^2 I.$$

Таким образом, матрица ковариаций не зависит от R.

Вернемся теперь к задаче оптимизации (2). Эту задачу можно эквивалентно переписать следующим образом:

$$\log p(X|W,\boldsymbol{\mu},\sigma) = \sum_{n=1}^{N} \log \mathcal{N}(\boldsymbol{x}_n|\boldsymbol{\mu},\sigma^2 I + WW^T) =$$

$$-\frac{ND}{2}\log 2\pi - \frac{N}{2}\det(\sigma^2 I + WW^T) - \frac{1}{2}\sum_{n=1}^{N}(\boldsymbol{x}_n-\boldsymbol{\mu})^T(\sigma^2 I + WW^T)^{-1}(\boldsymbol{x}_n-\boldsymbol{\mu}) \to \max_{W,\boldsymbol{\mu},\sigma}$$

Можно показать (см. [7]), что данная задача оптимизации имеет аналитическое решение:

$$\begin{aligned} \boldsymbol{\mu} &= \frac{1}{N}\sum_{n=1}^{N}\boldsymbol{x}_n, \\ W &= Q(\Lambda - \sigma^2 I)^{1/2}R, \\ \sigma^2 &= \frac{1}{D-d}\sum_{i=d+1}^{D}\lambda_i. \end{aligned} \tag{4}$$

Здесь $Q = (\boldsymbol{q}_1|\ldots|\boldsymbol{q}_d) \in \mathbb{R}^{D\times d}$, $\boldsymbol{q}_1,\ldots,\boldsymbol{q}_d$ – собственные вектора выборочной матрицы ковариации, отвечающие наибольшим собственным значениям $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_d$, $\Lambda = \operatorname{diag}(\lambda_1,\ldots,\lambda_d)$, R – произвольная ортогональная матрица размера $d \times d$.

Рассмотрим подробнее решение (4). Заметим, что в отличие от классической модели PCA, в которой восстанавливается только гиперплоскость, наилучшим образом объясняющая данные, вероятная модель PCA восстанавливает всю модель изменчивости данных и, в частности, дисперсии данных по всем направлениям. Поэтому решение (4) включает в себя не только направляющие базисные вектора гиперплоскости, задаваемые собственными векторами матрицы ковариаций, но также и длины этих базисных векторов, определяемые значениями $\sqrt{\lambda_i - \sigma^2}$ (в предположении, что дисперсии скрытых компонент равны единице). Действительно, величина дисперсии данных вдоль направления $\boldsymbol{v}$ составляет $\boldsymbol{v}^T C\boldsymbol{v}$, где $\boldsymbol{v}^T\boldsymbol{v} = 1$. Если $\boldsymbol{v}$ лежит в подпространстве, ортогональном гиперплоскости, то $\boldsymbol{v}^T C\boldsymbol{v} = \sigma^2$.

Теперь пусть $\boldsymbol{v}$ совпадает с одним из собственных векторов $\boldsymbol{q}_i$. Тогда $\boldsymbol{v}^T C\boldsymbol{v} = \lambda_i - \sigma^2 + \sigma^2 = \lambda_i$. Таким образом, вероятностная модель PCA корректно восстанавливает дисперсии данных в пространстве гиперплоскости и аппроксимирует дисперсию средним значением в ортогональном пространстве.

Зная параметры $W, \boldsymbol{\mu}, \sigma$, задача поиска для объекта $\boldsymbol{x}$ представления $\boldsymbol{t}$ в пространстве $\mathbb{R}^d$ сводится к вычислению математического ожидания условного распределения

$$p(\boldsymbol{t}|\boldsymbol{x}) = \frac{p(\boldsymbol{t},\boldsymbol{x})}{\int p(\boldsymbol{t},\boldsymbol{x})d\boldsymbol{x}} = \mathcal{N}(\boldsymbol{t}|(\sigma^2 I + W^T W)^{-1}W^T(\boldsymbol{x}-\boldsymbol{\mu}), I + \sigma^{-2}W^T W),$$
$$\mathbb{E}_{\boldsymbol{t}|\boldsymbol{x}}\boldsymbol{t} = (\sigma^2 I + W^T W)^{-1}W^T(\boldsymbol{x}-\boldsymbol{\mu}).$$

Как было отмечено выше, рассмотрение вероятностной модели со скрытыми переменными позволяет решать задачу максимизации правдоподобия с помощью итерационного EM-алгоритма. Однако, прежде чем перейти к рассмотрению EM-алгоритма для модели PPCA, рассмотрим схему EM-алгоритма в общем виде и ее применение для восстановления смеси нормальных распределений.

EM-алгоритм в общем виде

Пусть имеется вероятностная модель, задаваемая совместным распределением $p(X, T|\Theta)$. Здесь X – набор наблюдаемых переменных, T – набор ненаблюдаемых переменных и Θ – набор параметров модели. Задача состоит в оценке параметров модели Θ с помощью метода максимального правдоподобия:

$$\log p(X|\Theta) = \log \int p(X,T|\Theta)dT \to \max_{\Theta}. \tag{5}$$

EM-алгоритм для решения этой задачи представляет собой итерационную процедуру. Пусть фиксировано некоторое значение параметров Θ_{old}. На E-шаге алгоритма вычисляется распределение значений скрытых переменных при данных параметрах:

$$p(T|X,\Theta_{old}) = \frac{p(X,T|\Theta_{old})}{\int p(X,T|\Theta_{old})dT}.$$

Затем на M-шаге новое значение параметров находится с помощью максимизации полного правдоподобия, усредненного по апостериорному распределению для T:

$$\Theta = \arg\max_{\Theta} \mathbb{E}_{T|X,\Theta_{old}} \log p(X,T|\Theta). \tag{6}$$

Шаги E и M повторяются в цикле до сходимости. Можно показать, что в процессе EM-итераций значение правдоподобия $p(X|\Theta)$ не убывает. Таким образом, EM-алгоритм позволяет находить локальный максимум правдоподобия.

Заметим, что во многих практических случаях решение задачи (6) намного проще, чем решение задачи (5). В частности, во всех рассматриваемых ниже моделях задача оптимизации на M шаге может быть решена аналитически.

Вычисление значения функции правдоподобия $p(X|\Theta)$ в фиксированной точке Θ требует интегрирования по пространству T и в ряде случаев может представлять собой вычислительно трудоемкую задачу. Заметим, что эта величина правдоподобия стоит также в знаменателе формулы на E шаге. Однако, апостериорное распределение $p(T|X,\Theta_{old})$, вычисляемое на

E шаге, используется затем только для вычисления математического ожидания логарифма полного правдоподобия на M шаге. Как правило, здесь не требуется знать все апостериорное распределение целиком, а достаточно знать лишь несколько статистик этого распределения (например, только мат.ожидания отдельных компонент $\mathbb{E}_{T|X,\Theta_{old}} t_n$ и парные ковариации $\mathbb{E}_{T|X,\Theta_{old}} t_n t_k$). Поэтому EM-алгоритм может быть вычислительно эффективен даже в тех случаях, когда вычисление значения правдоподобия $p(X|\Theta)$ в одной точке затруднено.

EM-алгоритм для разделения гауссовской смеси

Рассмотрим вероятностную модель гауссовской смеси:

$$p(\boldsymbol{x}) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}_k, \Sigma_k), \;\; \sum_{k=1}^{K} \pi_k = 1, \;\; \pi_k \geq 0, \tag{7}$$

где $\Sigma_k : \Sigma_k = \Sigma_k^T, \Sigma_k \succ 0$ (символом $\Sigma \succ 0$ обозначены положительно определенные матрицы). Задача восстановления параметров смеси $\boldsymbol{\pi}, M = \{\boldsymbol{\mu}_k\}_{k=1}^K, S = \{\Sigma_k\}_{k=1}^K$ по выборке $X = \{\boldsymbol{x}_n\}_{n=1}^N$ с помощью метода максимального правдоподобия может быть записана как

$$\begin{cases} \log p(X|\boldsymbol{\pi}, M, S) = \sum_{n=1}^{N} \log\left(\sum_{k=1}^{K} \pi_k \mathcal{N}(\boldsymbol{x}_n|\boldsymbol{\mu}_k, \Sigma_k)\right) \to \max\limits_{\boldsymbol{\pi}, M, S}, \\ \sum_{k=1}^{K} \pi_k = 1, \;\; \pi_k \geq 0, \\ \Sigma_k = \Sigma_k^T, \;\; \Sigma_k \succ 0. \end{cases} \tag{8}$$

Построим вероятностную модель со скрытыми переменными, эквивалентную (7). Для этого введем бинарный вектор $\boldsymbol{t} \in \{0,1\}^K$ длины K, в котором только одно значение равно единице. Рассмотрим следующую вероятностную модель:

$$p(\boldsymbol{t}) = \prod_{k=1}^{K} \pi_k^{t_k}, \tag{9}$$

$$p(\boldsymbol{x}|\boldsymbol{t}) = \prod_{k=1}^{K} \left(\mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}_k, \Sigma_k)\right)^{t_k}. \tag{10}$$

Генерация объекта из модели (9)-(10) происходит следующим образом. Сначала с вероятностями, пропорциональными $\boldsymbol{\pi}$, генерируется номер компоненты смеси, из которой затем генерируется точка $\boldsymbol{x}$. Можно показать, что маргинальное распределение $p(\boldsymbol{x}) = \int p(\boldsymbol{x}|\boldsymbol{t}) p(\boldsymbol{t}) d\boldsymbol{t}$ в модели (9)-(10) совпадает с распределением (7). В этом смысле модели (9)-(10) и (7) эквивалентны.

Рассмотрим теперь применение EM-алгоритма для вероятностой модели со скрытыми переменными (9)-(10) для решения задачи максимизации правдоподобия (8). Для этого вычислим значение мат.ожидания логарифма полного правдоподобия, необходимого для решения задачи оптимизации на M шаге:

$$\mathbb{E} \log p(X, T|\boldsymbol{\pi}, M, S) = \sum_{n=1}^{N} \sum_{k=1}^{K} \mathbb{E} t_{nk} (\log \pi_k + \log \mathcal{N}(\boldsymbol{x}_n|\boldsymbol{\mu}_k, \Sigma_k)). \tag{11}$$

Заметим, что это выражение зависит только от мат.ожиданий отдельных компонент t_{nk}. Нетрудно показать, что эти величины можно вычислить следующим образом:

$$\gamma_{nk} = \mathbb{E}_{T|X,\boldsymbol{\pi}^{old},M^{old},S^{old}} t_{nk} = \frac{\pi_k^{old} \mathcal{N}(\boldsymbol{x}_n|\boldsymbol{\mu}_k^{old}, \Sigma_k^{old})}{\sum_{j=1}^{K} \pi_j^{old} \mathcal{N}(\boldsymbol{x}_n|\boldsymbol{\mu}_j^{old}, \Sigma_j^{old})}. \tag{12}$$

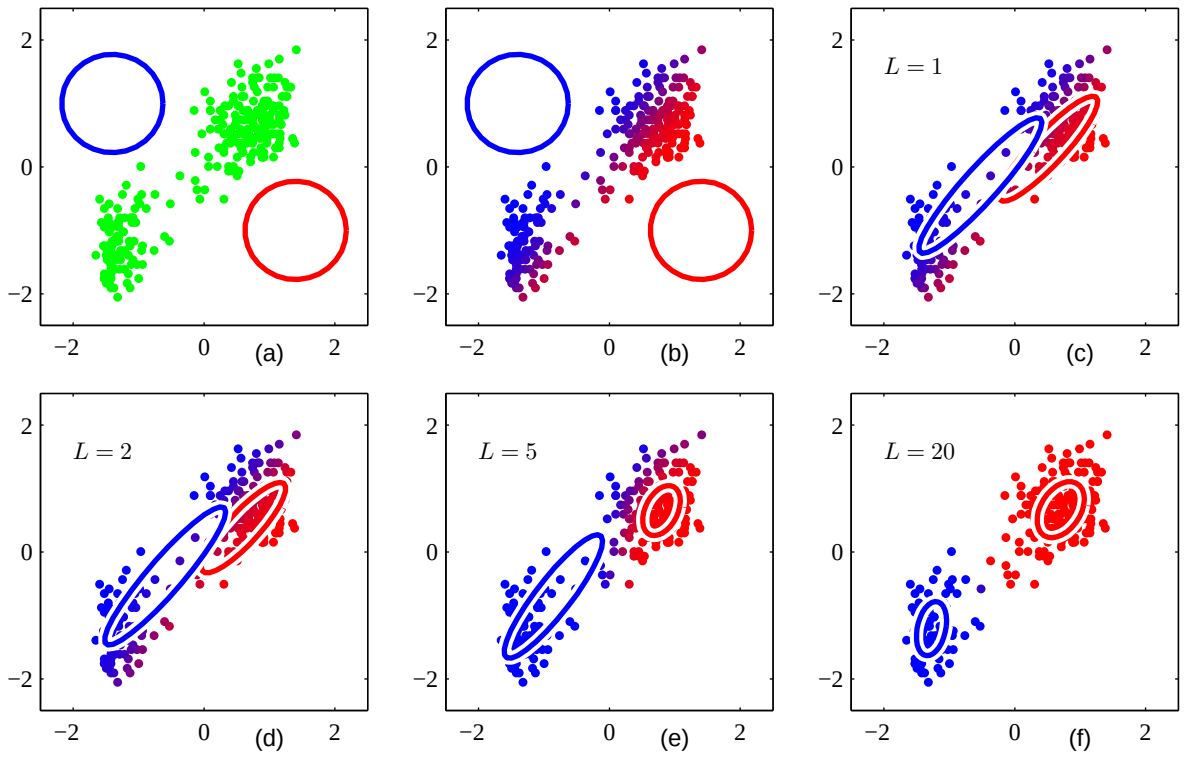


Рис. 6: Иллюстрация применения EM-алгоритма для разделения смеси нормальных распределений с двумя компонентами. На рис. a показана исходная выборка и начальное приближение для двух компонент. На рис. b показан результат E шага. При этом цвета объектов соответствуют значениям γ_{nk}. На рис. c-f показаны результаты вычислений после 1, 2, 5 и 20 итераций.

Также нетрудно показать, что задача максимизации критерия (11) при ограничениях $\sum_{k=1}^{K} \pi_k = 1,\ \pi_k \geq 0$ может быть решена аналитически:

$$\pi_k = \frac{1}{N} \sum_{n=1}^{N} \gamma_{nk}, \tag{13}$$

$$\boldsymbol{\mu}_k = \frac{\sum_{n=1}^{N} \gamma_{nk} \boldsymbol{x}_n}{\sum_{n=1}^{N} \gamma_{nk}}, \tag{14}$$

$$\Sigma_k = \frac{\sum_{n=1}^{N} \gamma_{nk} (\boldsymbol{x}_n - \boldsymbol{\mu}_k)(\boldsymbol{x}_n - \boldsymbol{\mu}_k)^T}{\sum_{n=1}^{N} \gamma_{nk}}. \tag{15}$$

Заметим, что решение для Σ_k (15) удовлетворяет условию симметричности и положительной определенности. Кроме того, формулы (14),(15) соответствуют оценке максимального правдоподобия для многомерного нормального распределения, в которых каждый объект $\boldsymbol{x}_n$ берется с весом γ_{nk}.

Таким образом, EM-алгоритм для смеси нормальных распределений заключается в итерационном применении формул (12) и (13)-(15). Этот процесс имеет простую интерпретацию. Величина γ_{nk} показывает степень соответствия между объектом $\boldsymbol{x}_n$ и компонентой k (определяет вес объекта $\boldsymbol{x}_n$ для компоненты k). Эти веса затем используются на M шаге для

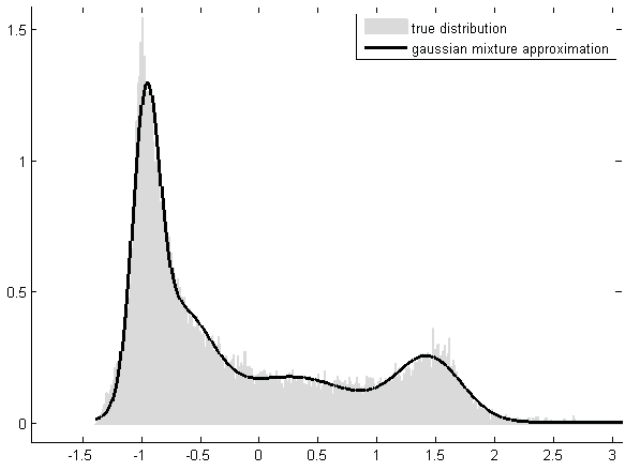


Рис. 7: Пример реального распределения (оцененного с помощью гистограммы) и его аппроксимация с помощью смеси пяти гауссиан.

вычисления новых значений параметров компонент. Иллюстрация применения EM-алгоритма для разделения нормальной смеси с двумя компонентами показана на рис. 6.

Одно из применений смеси нормальных распределений – аналитическая аппроксимация сложных распределений, возникающих на практике. На рис. 7 приведен пример распределения выборки (оцененный с помощью гистограммы) и его аппроксимация с помощью смеси пяти гауссиан.

Другим применением смеси нормальных распределений является решение задачи кластеризации на K кластеров. В этом случае номер кластера для объекта $\boldsymbol{x}_n$ определяется величиной

$$k_n = \arg\max_k \gamma_{nk}.$$

Такая схема кластеризации является вероятностным обобщением известного метода кластеризации K-средних.

EM-алгоритм для вероятностного метода главных компонент

Рассмотрим теперь применение EM-алгоритма для вероятностной модели PCA:

$$p(X, T|W, \sigma^2, \boldsymbol{\mu}) = \prod_{n=1}^{N} p(\boldsymbol{x}_n|\boldsymbol{t}_n, W, \boldsymbol{\mu}, \sigma^2)p(\boldsymbol{t}_n),$$
$$p(\boldsymbol{x}_n|\boldsymbol{t}_n, W, \boldsymbol{\mu}, \sigma^2) = \mathcal{N}(\boldsymbol{x}_n|W\boldsymbol{t}_n + \boldsymbol{\mu}, \sigma^2 I),$$
$$p(\boldsymbol{t}_n) = \mathcal{N}(\boldsymbol{t}_n|\boldsymbol{0}, I).$$

E-шаг:

$$p(T|X, W_{old}, \sigma^2_{old}, \boldsymbol{\mu}_{old}) = \prod_{n=1}^{N} p(\boldsymbol{t}_n|\boldsymbol{x}_n, W_{old}, \sigma^2_{old}, \boldsymbol{\mu}_{old}),$$
$$p(\boldsymbol{t}_n|\boldsymbol{x}_n, W_{old}, \sigma^2_{old}, \boldsymbol{\mu}_{old}) = \mathcal{N}(\boldsymbol{t}_n|\boldsymbol{\mu}_n, \Sigma_n),$$
$$\boldsymbol{\mu}_n = M_{old} W^T_{old}(\boldsymbol{x}_n - \boldsymbol{\mu}_{old}),$$
$$\Sigma_n = \sigma^2_{old} M_{old},$$
$$M_{old} = (W^T_{old} W_{old} + \sigma^2_{old} I)^{-1}.$$

M-шаг:

$$\boldsymbol{\mu}_{new} = \frac{1}{N}\sum_{n=1}^{N} \boldsymbol{x}_n,$$
$$W_{new} = \left(\sum_{n=1}^{N}(\boldsymbol{x}_n - \boldsymbol{\mu}_{new})\mathbb{E}\boldsymbol{t}^T_n\right)\left(\sum_{n=1}^{N}\mathbb{E}\boldsymbol{t}_n\boldsymbol{t}^T_n\right)^{-1},$$
$$\sigma^2_{new} = \frac{1}{ND}\sum_{n=1}^{N}\Big((\boldsymbol{x}_n - \boldsymbol{\mu}_{new})^T(\boldsymbol{x}_n - \boldsymbol{\mu}_{new}) - 2\mathbb{E}\boldsymbol{t}^T_n W^T_{new}(\boldsymbol{x}_n - \boldsymbol{\mu}_{new}) + \mathrm{tr} W^T_{new} W_{new}\mathbb{E}\boldsymbol{t}_n\boldsymbol{t}^T_n\Big).$$

При этом необходимые статистики вычисляются следующим образом:

$$\mathbb{E}\boldsymbol{t}_n = \boldsymbol{\mu}_n,$$
$$\mathbb{E}\boldsymbol{t}_n\boldsymbol{t}^T_n = \Sigma_n + \boldsymbol{\mu}_n\boldsymbol{\mu}^T_n.$$

Заметим, что в процессе EM-итераций значение $\boldsymbol{\mu}$ не меняется и равно выборочному среднему. Поэтому при практическом применении EM-алгоритма для PPCA выборка X сначала центрируется, а затем все вычисления проводятся без пересчета $\boldsymbol{\mu}$.

EM-алгоритм сходится к решению для W (4) с некоторой неединичной матрицей R. Таким образом, полученные признаки не обладают свойством некоррелированности. Кроме того, столбцы W не являются, вообще говоря, ортогональными. Поэтому для получения ортогонального базиса требуется дополнительно проводить процесс ортогонализации Грамма-Шмидта.

Как уже было отмечено выше, итерационный EM-алгоритм является вычислительно более эффективным по сравнению с аналитическим решением (4) для больших выборок и в ситуациях, когда $d \ll D$. Действительно, вычисление выборочной матрицы ковариаций требует $O(ND^2)$, а поиск ее собственных значений и собственных векторов – $O(D^3)$ или $O(N^3)$. В EM-алгоритме самые сложные операции требуют $O(NDd)$ и $O(d^3)$, что может дать существенный выигрыш при больших N и $d \ll D$.

2.1 Учет пропусков в данных

Одно из преимуществ использования EM-алгоритма для максимизации правдоподобия в модели PPCA – возможность прямого обобщения метода на случай наличия пропусков в данных. Обозначим K_n — множество известных значений объекта $\boldsymbol{x}_n$ и U_n — множество пропущенных значений объекта $\boldsymbol{x}_n$, $K_n \cup U_n = \{1, \ldots, D\}$. Соответственно $W_{K_n} = \{w_{ij}\}_{i\in K_n, j\in\{1,\ldots,d\}}$.

Вероятностная модель PPCA с пропусками в данных выглядит следующим образом:

$$p(X_K, X_U, T|W, \sigma^2, \boldsymbol{\mu}) = \prod_{n=1}^{N} p(\boldsymbol{x}_{n,K_n}, \boldsymbol{x}_{n,U_n}|\boldsymbol{t}_n, W, \boldsymbol{\mu}, \sigma^2)p(\boldsymbol{t}_n),$$
$$p(\boldsymbol{x}_{n,K_n}, \boldsymbol{x}_{n,U_n}|\boldsymbol{t}_n, W, \boldsymbol{\mu}, \sigma^2) = \mathcal{N}((\boldsymbol{x}_{n,K_n}, \boldsymbol{x}_{n,U_n})|(W_{K_n}\boldsymbol{t}_n + \boldsymbol{\mu}_{K_n}, W_{U_n}\boldsymbol{t}_n + \boldsymbol{\mu}_{U_n}), \sigma^2 I),$$
$$p(\boldsymbol{t}_n) = \mathcal{N}(\boldsymbol{t}_n|\boldsymbol{0}, I).$$

Нетрудно показать, что EM-алгоритм для этой модели состоит в следующем:
E-шаг:

$$p(X_U, T|X_K, W_{old}, \sigma^2_{old}) = \prod_{n=1}^{N} p((\boldsymbol{x}_{n,U_n}, \boldsymbol{t}_n)|\boldsymbol{x}_{n,K_n}, W_{old}, \sigma^2_{old}),$$
$$p((\boldsymbol{x}_{n,U_n}, \boldsymbol{t}_n)|\boldsymbol{x}_{n,K_n}, W_{old}, \sigma^2_{old}) = \mathcal{N}\Big((\boldsymbol{x}_{n,U_n}, \boldsymbol{t}_n)|\boldsymbol{m}_n, S_n\Big),$$
$$\boldsymbol{m}_n = (W_{U_n}MW^T_{K_n}\boldsymbol{x}_{n,K_n}, MW^T_{K_n}\boldsymbol{x}_{n,K_n}),$$
$$S_n = \sigma^2_{old}\begin{pmatrix} I + W_{U_n}MW^T_{U_n} & -W_{U_n}M \\ -MW^T_{U_n} & M \end{pmatrix},$$
$$M = (W^T_{K_n}W_{K_n} + \sigma^2_{old}I)^{-1}.$$

M-шаг:

$$W^{new}_i = \left(\sum_{n:i\in K_n} x_{ni}\mathbb{E}\boldsymbol{t}^T_n + \sum_{n:i\in U_n}\mathbb{E}x_{ni}\boldsymbol{t}^T_n\right)\left(\sum_{n=1}^{N}\mathbb{E}\boldsymbol{t}_n\boldsymbol{t}^T_n\right)^{-1},$$
$$\sigma^2_{new} = \frac{1}{ND}\sum_{n=1}^{N}\Big(\boldsymbol{x}^T_{n,K_n}\boldsymbol{x}_{n,K_n} + \mathrm{tr}\mathbb{E}\boldsymbol{x}_{n,U_n}\boldsymbol{x}^T_{n,U_n} - 2\mathbb{E}\boldsymbol{t}^T_n W^T_{K_n}\boldsymbol{x}_{n,K_n} - 2\mathrm{tr}W^T_{U_n}\mathbb{E}\boldsymbol{x}_{n,U_n}\boldsymbol{t}^T_n +$$
$$+ \mathrm{tr}W^T_{K_n}W_{K_n}\mathbb{E}\boldsymbol{t}_n\boldsymbol{t}^T_n + \mathrm{tr}W^T_{U_n}W_{U_n}\mathbb{E}\boldsymbol{t}_n\boldsymbol{t}^T_n\Big).$$

При этом выборка предварительно центрируется на величину

$$\boldsymbol{\mu}_i = \frac{\sum_{n:i\in K_n}^{N}\boldsymbol{x}_{ni}}{\sum_{n:i\in K_n}1}.$$

Заметим, что формулы EM-алгоритма для модели PPCA с пропусками переходят в соответствующие формулы EM-алгоритма для PPCA в том случае, если пропусков в данных нет.

В качестве иллюстративного примера вернемся к задаче выбора признакового пространства для базы данных рукописных цифр MNIST. На рис. 8 приведена проекция исходной выборки на первые две главные компоненты (совпадает с рис. 2b), а также аналогичная проекция для выборки, в которой 30% случайно выбранных значений считаются пропущенными. Как видно, результаты практически совпадают между собой.

Рассмотренный метод учета пропусков в данных является адекватным для случая, когда места пропусков в данных определяются случайными факторами и, в частности, не зависят от истинных значений признаков в местах пропуска. Если, например, измерительный датчик дает сбой только для экстремальных значений измеряемой характеристики, то здесь необходима модификация вероятностной модели с пропусками, учитывающей модель образования пропущенных значений.

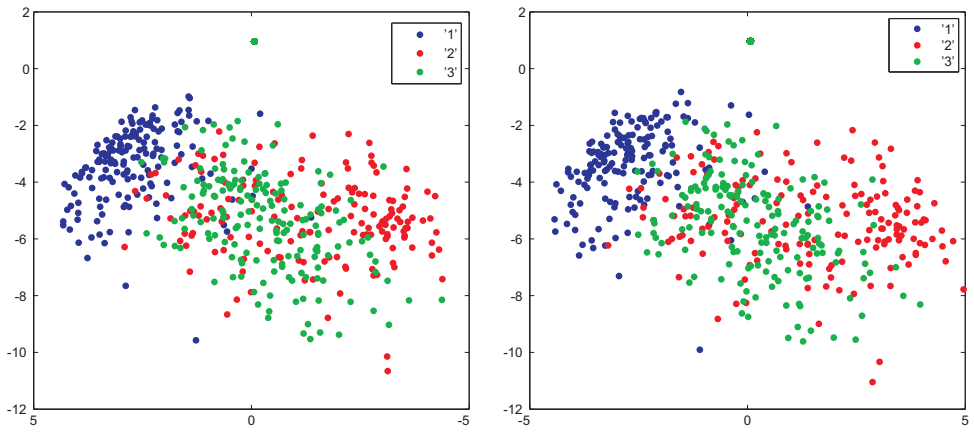


Рис. 8: Проекция выборки изображений цифр ’1’, ’2’, ’3’ на первые две главные компоненты для полных данных (a) и для выборки, в которой 30% случайно выбранных значений считаются пропущенными.

3 Обобщения метода главных компонент

Отметим ряд ограничений метода главных компонент. Первое ограничение связано с линейностью метода. В том случае, если выборка данных образует скрытую поверхность, которая является существенно нелинейной, метод главных компонент может приводить к неадекватным результатам (большая ошибка при восстановлении данных или маленькая редукция размерности пространства). Простым обобщением метода главных компонент, которое позволяет преодолеть это ограничение, является рассмотрение вероятностной смеси главных компонент.

Другим ограничением метода главных компонент является произвол в выборе базиса в пространстве оптимальной гиперплоскости, т.е. возможность определения скрытых компонент T только с точностью до афинного преобразования. В ряде случаев подобный произвол может приводить к неадекватным результатам. Рассмотрим здесь в качестве примера задачу разделения независимых источников. Пусть имеется набор одномерных сигналов $t^i(l_n)$, заданных в моменты времени $l_1, \ldots, l_N$. При этом наблюдаются не исходные сигналы, а совокупность их линейных комбинаций, т.е. набор сигналов $x^j(l_n)$, где

$$x^j(l_n) = \sum_i w_{ij} t^i(l_n).$$

Здесь $w_{ij} \in \mathbb{R}$ – некоторые веса, а количество сигналов X может быть как больше, так и меньше, чем число исходных сигналов T. Задача состоит в том, чтобы по набору наблюдаемых сигналов X восстановить исходные сигналы T. Практическим примером данной задачи является задача построения магнитоэнцефалограммы головного мозга по набору датчиков. Каждый датчик выдает сигнал, который является комбинацией магнитной активности в данной области мозга, а также части магнитной активности с соседних датчиков. В результате возникает необходимость выделения собственных сигналов для каждого датчика. Другим практическим примером является восстановление исходных дорожек в музыкальном стерео-сигнале.

Будем решать задачу разделения независимых источников с помощью метода главных компонент следующим образом. Будем трактовать наблюдения сигналов как совокупность

независимых точек, т.е. рассмотрим выборку из N объектов, где n-ый объект состоит из $x^1(l_n), x^2(l_n), \ldots, x^D(l_n)$. С помощью метода уменьшения размерности найдем совокупность N объектов вида $t^1(l_n), t^2(l_n), \ldots, t^d(l_n)$. Как уже было отмечено выше, при использовании метода главных компонент результат является произвольным с точностью до афинного преобразования. Очевидно, что такой подход, вообще говоря, не позволяет выявить исходные компоненты сигнала. Для решения данной задачи можно воспользоваться подходом «Анализ независимых факторов» (см. [8]), который является обобщением вероятностного метода главных компонент.

3.1 Вероятностная смесь главных компонент

Рассмотрим следующую вероятностную модель:

$$p(\boldsymbol{x}) = \sum_{k=1}^{K} \pi_k p_k(\boldsymbol{x}|W_k, \sigma_k^2, \boldsymbol{\mu}_k) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}_k, W_k W_k^T + \sigma_k^2 I),\ \sum_{k=1}^{K} \pi_k = 1,\ \pi_k \geq 0.$$

Эта модель представляет собой смесь нормальных распределений, в которой матрицы ковариации задаются специальным образом. Введем эквивалентную вероятностную модель путем добавления скрытых переменных $\boldsymbol{z}_n \in \{0,1\}^K, \sum_{k=1}^K z_{nk} = 1$ для каждого объекта $\boldsymbol{x}_n$, отвечающих за номер компоненты смеси:

$$p(\boldsymbol{z}_n) = \prod_{k=1}^{K} \pi_k^{z_{nk}},$$

$$p(\boldsymbol{x}_n|\boldsymbol{z}_n) = \prod_{k=1}^{K} (p_k(\boldsymbol{x}_n))^{z_{nk}}.$$

Можно показать (см. [7]), что EM-алгоритм максимизации правдоподобия в этой модели по параметрам $\boldsymbol{\pi}$, $M = \{\boldsymbol{\mu}_k\}_{k=1}^K$, $\mathcal{W} = \{W_k\}_{k=1}^K$ и $\boldsymbol{\sigma}$ выглядит следующим образом:
E-шаг:

$$\gamma_{nk} = \mathbb{E}_{Z|X,\boldsymbol{\pi},M,\mathcal{W},\boldsymbol{\sigma}} z_{nk} = \frac{\pi_k p_k(\boldsymbol{x}_n|W_k, \sigma_k^2, \boldsymbol{\mu}_k)}{\sum_{j=1}^K \pi_j p_j(\boldsymbol{x}_n|W_j, \sigma_j^2, \boldsymbol{\mu}_j)}.$$

M-шаг:

$$\pi_k^{new} = \frac{1}{N} \sum_{n=1}^{N} \gamma_{nk},$$

$$\boldsymbol{\mu}_k^{new} = \frac{\sum_{n=1}^N \gamma_{nk} \boldsymbol{x}_n}{\sum_{n=1}^N \gamma_{nk}},$$

$$S_k = \frac{\sum_{n=1}^N \gamma_{nk} (\boldsymbol{x}_n - \boldsymbol{\mu}_k)(\boldsymbol{x}_n - \boldsymbol{\mu}_k)^T}{\sum_{n=1}^N \gamma_{nk}}.$$

При этом параметры W_k и σ_k вычисляются по стандартным формулам для PPCA путем разложения по собственным векторам матрицы S_k. Альтернативные формулы пересчета этих параметров без привлечения промежуточной матрицы ковариации S_k можно найти в [7]. Эти формулы также можно легко обобщить на случай наличия пропусков в данных.

Заметим, что формулы для γ_{nk}, π_k и $\boldsymbol{\mu}_k$ полностью совпадают с соответствующими формулами EM-алгоритма для смеси нормальных распределений, рассмотренных выше.

Восстановление вероятностной смеси главных компонент соответствует построению K линейных подпространств, определяемых параметрами $W_k, \boldsymbol{\mu}_k, \sigma_k$. Таким образом, для заданного объекта $\boldsymbol{x}_n$ можно получить проекцию $\boldsymbol{t}_n$ на подпространство с номером k по формуле $(\sigma_k^2 I + W_k^T W_k)^{-1} W_k^T (\boldsymbol{x}_n - \boldsymbol{\mu}_k)$. При этом по аналогии с применением смеси нормальных распределений для решения задачи кластеризации, номер подпространства k выбирается как

$$k = \arg\max_k \gamma_{nk}. \tag{16}$$

С точки зрения задачи уменьшения размерности в данных, для каждого объекта $\boldsymbol{x}_n$ сохраняется номер подпространства k и проекция объекта на это подпространство $\boldsymbol{t}_n$. Заметим, что выбор подпространства с помощью формулы (16), вообще говоря, не соответствует выбору подпространства с наименьшей квадратичной ошибкой восстановления.

Применение вероятностной смеси главных компонент

Модель смеси главных компонент имеет широкую область применения. Помимо решения задачи уменьшения размерности и сжатия данных, эту модель можно использовать для решения задачи кластеризации и для восстановления плотности распределения выборки.

Рассмотрим применение модели смеси главных компонент для задачи кластеризации. Как уже было отмечено выше, эта модель является частным случаем общей модели смеси нормальных распределений, в которой матрица ковариации для каждой компоненты задается специальным образом: $C_k = W_k W_k^T + \sigma_k^2 I$. Как и в общем случае, для решения задачи кластеризации на K кластеров сначала восстанавливаются параметры модели смеси $\{W_k, \boldsymbol{\mu}_k, \sigma_k\}_{k=1}^K$ с помощью описанного выше EM-алгоритма, а затем номер кластера для объекта $\boldsymbol{x}_n$ определяется с помощью формулы (16).

Матрица ковариации в модели смеси главных компонент требует задания $dD+1-d(d+1)/2$ параметров (величина $d(d+1)/2$ вычитается, т.к. матрица ковариации определяется с точностью до ортогональной матрицы поворота системы координат в линейном подпространстве). Произвольная симметричная неотрицательно определенная матрица размера $D \times D$ задается $D(D+1)/2$ параметрами, а диагональная матрица — D параметрами. Таким образом, модель смеси главных компонент имеет смысл применять для решения задачи кластеризации в том случае, когда восстановление смеси произвольных нормальных распределений не представляется возможным в силу ограниченности выборки (для полной смеси нужно определить $K(D(D+1)/2 + D + 1)$ параметров, а для смеси главных компонент только $K(dD + 1 - d(d+1)/2 + D + 1)$ параметров). Кроме того, при применении смеси произвольных нормальных распределений кластеры представляют собой компактные шарообразные формы, а то время как в смеси главных компонент кластеры образуют объекты, лежащие в одном линейном подпространстве заданной размерности.

В качестве примера рассмотрим кластеризацию двухмерной выборки данных, представленной на рис. 9a, с помощью трех методов: 1) восстановление смеси произвольных нормальных распределений (см. рис. 9b), 2) восстановление смеси главных компонент с $d = 1$ (см. рис. 9c) и 3) восстановление смеси диагональных нормальных распределений (см. рис. 9d). Под диагональным нормальным распределением понимается распределение с диагональной матрицей ковариации. Как видно из рисунков, смесь диагональных нормальных распределений показывает не совсем адекватный результат кластеризации, т.к. выборка не образует фрагментов, распределенных вдоль координатных осей. Смесь главных компонент кластеризует данные так, чтобы каждый кластер был максимально похож на прямую линию.

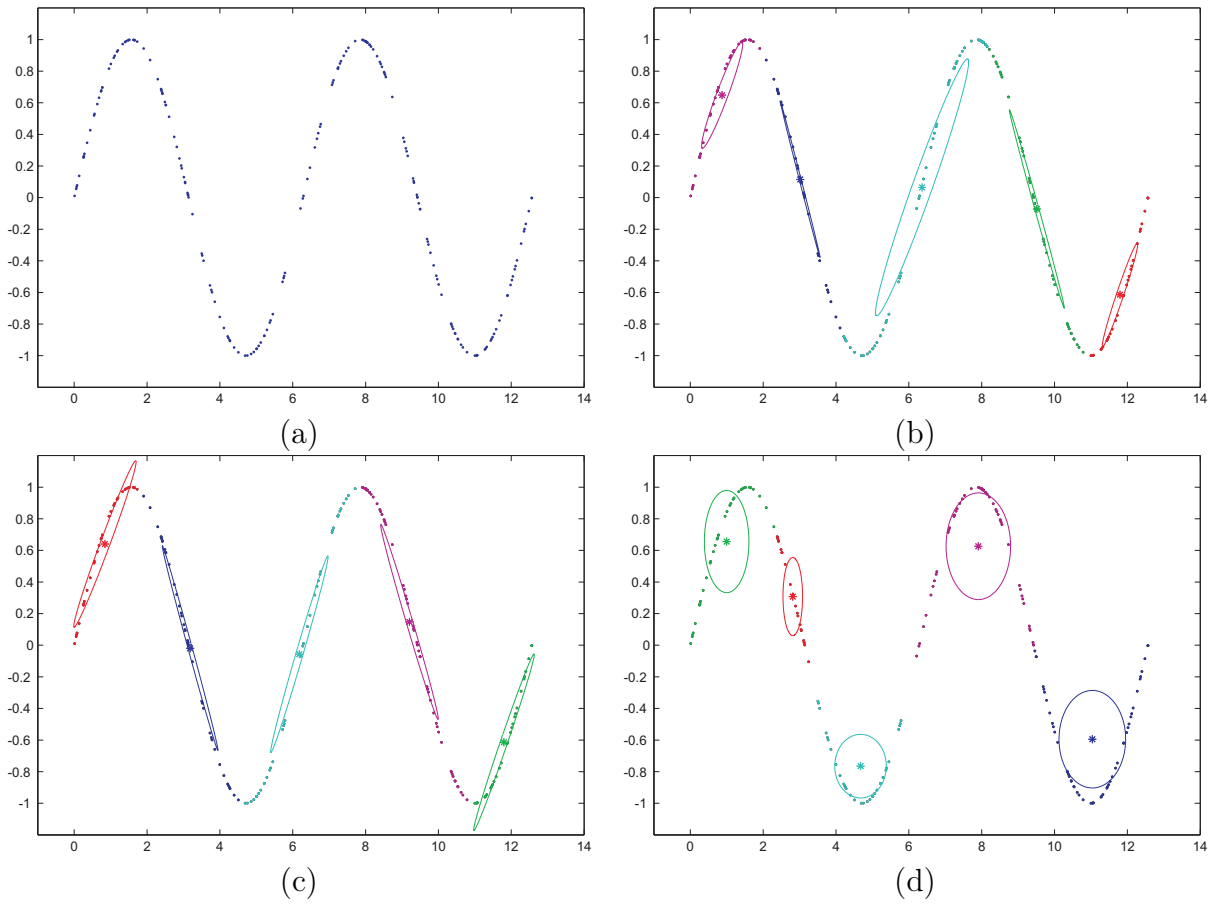
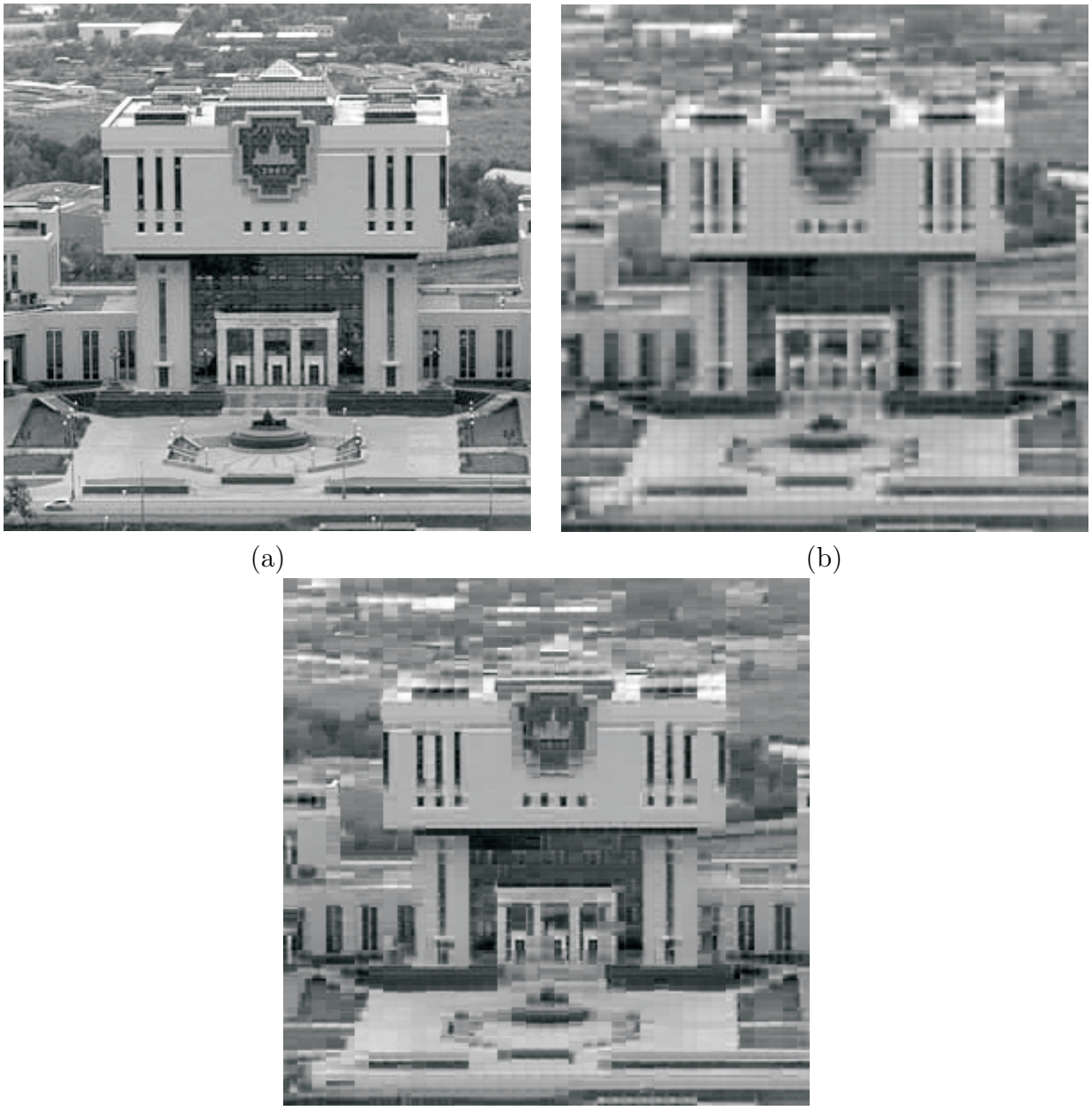


Рис. 9: Кластеризация двухмерной выборки (рис. a) на 5 кластеров с помощью трех методов: смесь произвольных нормальных распределений (рис. b), смесь главных компонент (рис. c) и смесь диагональных нормальных распределений (рис. d). Цветами обозначены объекты соответствующих кластеров. Кроме того, показаны центры и эллипсы рассеивания для каждой компоненты смеси.

Смесь произвольных нормальных распределений кластеризует данные похожим образом, однако средний кластер имеет отклонения от прямой линии, т.к. на концах кластера имеется компактная группа объектов.

Другим возможным применением модели смеси главных компонент является блочное сжатие изображений. Пусть имеется некоторое черно-белое изображение (см. рис. 10a). Разобьем это изображение на набор непересекающихся блоков размера 8×8, и каждый блок вытянем в вектор длины 64. Таким образом, мы получим некоторую выборку размера $<$Число_блоков$> \times 64$. Например, для изображения размера 304×200 соответствующая выборка будет иметь размер 950×64. Затем применим к этой выборке методы уменьшения размерности в данных для решения задачи сжатия изображения. На рис. 10 приведен пример применения метода блочного сжатия изображения с помощью вероятностного метода главных компонент с $d = 4$ (см. рис. 10b) и вероятностной смеси главных компонент с $d = 3$ и $K = 15$ (см. рис. 10c). В обоих случаях коэффициент сжатия равен 16 (в вероятностной смеси главных компонент помимо проекции на подпространство сохраняется дополнительно номер этого подпространства). Как видно из рисунка, смесь главных компонент обеспечивает меньшую величину ошибки и, соответственно, более высокое качество восстановления изображения.



(a) (b)

(c)

Рис. 10: Иллюстрация сжатия изображения (рис. a) в 16 раз с помощью метода главных компонент (рис. b) и смеси главных компонент (рис. c).

Стоит отдельно подчеркнуть, что рассмотренный метод является скорее иллюстративным примером к вероятностной модели смеси главных компонент, чем реальным методом сжатия изображений, т.к., например, он никак не учитывает специфику предметной области и особенности реалистичных изображений.

Еще одним примером применения модели смеси главных компонент является восстановление плотности классов при решении задачи классификации. Пусть имеется задача классификации на K классов. Восстановим по обучающей выборке плотность каждого из классов $p(\boldsymbol{x}|k)$ с помощью вероятностной модели смеси главных компонент. После этого можно воспользоваться

байесовским классификатором и классифицировать объекты по следующему правилу:

$$\hat{k}(\boldsymbol{x}) = \arg\max_k p(k|\boldsymbol{x}) = \arg\max_k p(\boldsymbol{x}|k)p(k).$$

Здесь $p(k)$ – априорная вероятность появления класса k. Заметим, что решение задачи классификации с помощью восстановления плотности каждого из классов требует большого объема обучающей выборки. Как уже было отмечено выше, модель смеси главных компонент задается значительно меньшим числом параметров, чем модель смеси произвольных нормальных распределений. В результате для применения модели смеси главных компонент требуется меньший объем обучающей выборки.

3.2 Анализ независимых факторов

Рассмотрим задачу разделения независимых источников, описанную выше. Как уже было отмечено ранее, метод главных компонент не подходит для решения этой задачи. В вероятностной формулировке метода в пространстве скрытых переменных $\boldsymbol{t}$ предполагается изотропное нормальное распределение $\mathcal{N}(\boldsymbol{t}|0, I)$. Можно утверждать, что нормальность распределения в пространстве скрытых переменных является ключевым моментом, не позволяющим решать задачу разделения независимых источников. Одним из основных свойств нормального распределения является тот факт, что его любые маргинальные распределения, а также распределение любой линейной комбинации переменных, являются нормальными. Пусть $p(x_1, x_2)$ является двумерным нормальным распределением. Тогда $p(x_1), p(x_2), p(x_1|x_2), p(x_2|x_1), p(\alpha_1 x_1 + \alpha_2 x_2)$ тоже являются нормальными. Таким образом, если исходные сигналы T имеют нормальное распределение, то нет никакой возможности их найти по наблюдаемым линейным комбинациям X. Предположим далее, что выборка является центрированной, т.е. $\sum_{n=1}^{N} \boldsymbol{x}_n = \mathbf{0}$ и $d = D$. Обозначим через $\boldsymbol{t}_{true}, W_{true}$ истинные исходные сигналы и матрицу смешивания. Прогноз скрытой переменной с помощью вероятностного метода главных компонент вычисляется как

$$\boldsymbol{t}_{PCA} = (W^T W)^{-1} W^T \boldsymbol{x} = (W^T W)^{-1} W^T W_{true} \boldsymbol{t}_{true}.$$

Если матрица смешивания W совпадает с истинной W_{true}, то тогда $\boldsymbol{t}_{PCA} = \boldsymbol{t}_{true}$. Если матрицы смешивания разные, то тогда прогноз t_{PCA} представляет собой линейную комбинацию независимых компонент $\boldsymbol{t}_{true}$ с весами, определяемыми матрицей $(W^T W)^{-1} W^T W_{true}$. По центральной предельной теореме сумма независимых случайных величин с ограниченными моментами стремится к нормальному распределению. Таким образом, практически при любых значениях W величина $\boldsymbol{t}_{PCA}$ стремится к нормальному распределению. Это означает, что вероятностная модель главных компонент всегда сможет найти такую матрицу W, чтобы распределение в пространстве $\boldsymbol{t}$ удовлетворяло бы предположениям модели, т.е. было бы нормальным. Однако, этот результат будет далек от исходных сигналов T. Таким образом, для решения задачи разделения независимых источников необходим отказ от предположения нормальности в пространстве скрытых переменных.

Рассмотрим модель «анализ независимых факторов». В этой модели, как и ранее, предполагается, что наблюдаемые переменные представляют собой зашумленную линейную комбинацию искомых компонент:

$$p(\boldsymbol{x}|\boldsymbol{t}) = \mathcal{N}(\boldsymbol{x}|W\boldsymbol{t}, \Lambda).$$

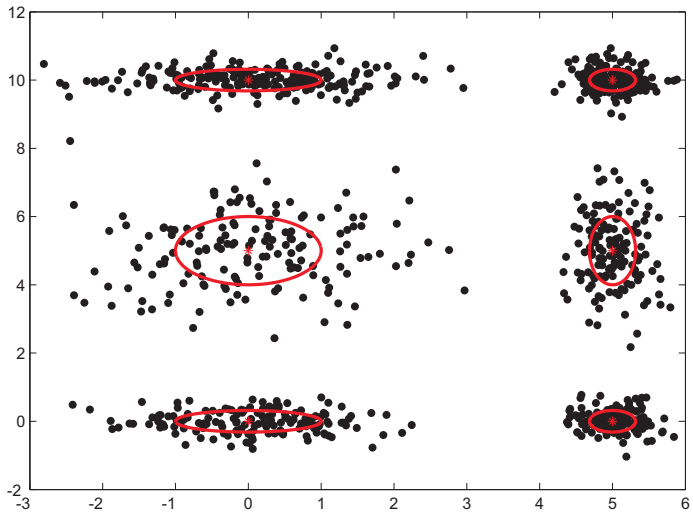


Рис. 11: Иллюстрация смеси нормальных распределений для модели «анализ независимых факторов». Здесь $d = 2, K_1 = 2, K_2 = 3$, $\mu_1^1 = 0, \mu_2^1 = 5$, $s_1^1 = 1, s_1^1 = 0.1$, $\mu_1^2 = 0, \mu_2^2 = 5, \mu_3^2 = 10$, $s_1^2 = 0.1, s_2^2 = 1, s_3^2 = 0.1$.

В отличие от модели главных компонент, здесь матрица ковариации шума Λ предполагается произвольной симметричной положительно определенной. Требование независимости скрытых источников (факторов) может быть выражено как

$$p(\boldsymbol{t}) = p(t^1)p(t^2)\dots p(t^d). \tag{17}$$

В модели главных компонент $p(t^i) = \mathcal{N}(t^i|0,1)$. В модели «анализ независимых факторов» $p(t^i)$ выбирается в параметрическом семействе распределений, способном приблизить с высокой точностью достаточно широкий класс непрерывных распределений, а именно в семействе смеси одномерных гауссиан:

$$p(t^i) = \sum_{j=1}^{K_i} \pi_j^i \mathcal{N}(t^i|\mu_j^i, s_j^i), \ \sum_{j=1}^{K_i} \pi_j^i = 1, \pi_j^i \geq 0. \tag{18}$$

Здесь K_i – количество компонент смеси для i-ого фактора, s_j^i – дисперсия j-ой компоненты для i-го фактора. Заметим, что если $K_i = 1 \ \forall i$, то модель «анализ независимых факторов» становится практически эквивалентной модели главных компонент. Подставляя (18) в (17), получаем, что распределение $p(\boldsymbol{t})$ в свою очередь есть смесь нормальных распределений:

$$p(\boldsymbol{t}) = \sum_{j_1,\dots,j_d=1}^{K_1,\dots,K_d} \pi_{j_1}^1 \dots \pi_{j_d}^d \mathcal{N}(\boldsymbol{t}|(\mu_{j_1}^1, \dots, \mu_{j_d}^d), \mathrm{diag}(s_{j_1}^1, \dots, s_{j_d}^d)). \tag{19}$$

Данная смесь нормальных распределений имеет сеточную структуру (см. рис. 11). Изменение параметров μ_j^i, s_j^i для одной компоненты распределения $p(t^i)$ влечет за собой изменение всех компонент распределения $p(\boldsymbol{t})$ в одной колонке.

Как и ранее при работе со смесями распределений введем набор вспомогательных переменных $\boldsymbol{z} = (z_1, \dots, z_d)$, где $z_i \in \{1, \dots, K_i\}$ – номер соответствующей компоненты смеси.

Тогда модель (19) можно эквивалентно переписать следующим образом:

$$p(\boldsymbol{t}|\boldsymbol{z}) = \mathcal{N}(\boldsymbol{t}|\boldsymbol{\mu}_{\boldsymbol{z}}, V_{\boldsymbol{z}}),\ \boldsymbol{\mu}_{\boldsymbol{z}} = (\mu^1_{z_1}, \ldots, \mu^d_{z_d}),\ V_{\boldsymbol{z}} = \operatorname{diag}(s^1_{z_1}, \ldots, s^d_{z_d}),$$
$$p(\boldsymbol{z}) = \pi^1_{z_1}\pi^2_{z_2}\ldots\pi^d_{z_d}.$$

Объединяя все вышесказанное, сформулируем модель «анализ независимых факторов»:

$$p(X, T, Z|W, M, S, \boldsymbol{\pi}) = \prod_{n=1}^{N} p(\boldsymbol{x}_n|\boldsymbol{t}_n)p(\boldsymbol{t}_n|\boldsymbol{z}_n)p(\boldsymbol{z}_n),$$
$$p(\boldsymbol{x}_n|\boldsymbol{t}_n) = \mathcal{N}(\boldsymbol{x}_n|W\boldsymbol{t}_n, \Lambda),$$
$$p(\boldsymbol{t}_n|\boldsymbol{z}_n) = \mathcal{N}(\boldsymbol{t}_n|\boldsymbol{\mu}_{\boldsymbol{z}_n}, V_{\boldsymbol{z}_n}),\ \boldsymbol{\mu}_{\boldsymbol{z}_n} = (\mu^1_{z_{n1}}, \ldots, \mu^d_{z_{nd}}), V_{\boldsymbol{z}_n} = \operatorname{diag}(s^1_{z_{n1}}, \ldots, s^d_{z_{nd}}),$$
$$p(\boldsymbol{z}_n) = \pi^1_{z_{n1}}\ldots\pi^d_{z_{nd}}.$$

Здесь переменные X являются наблюдаемыми, переменные (T, Z) – ненаблюдаемыми, $(W, M, S, \boldsymbol{\pi})$ – набор параметров, где $M = \{\mu^i_j\}^{d,K_i}_{i,j=1}$, $S = \{s^i_j\}^{d,K_i}_{i,j=1}$. Генерация объекта $\boldsymbol{x}$ из этой модели происходит в три этапа. Сначала с вероятностями, пропорциональными $\pi^1_{z_1}\ldots\pi^d_{z_d}$, генерируются номера компонент смеси $z_1, \ldots, z_d$ для каждого признака. Затем переменная $\boldsymbol{t}$ генерируется из нормального распределения, параметры которого задаются $\boldsymbol{z}$, а объект $\boldsymbol{x}$ в свою очередь генерируется как зашумленная линейная комбинация $W\boldsymbol{t}$.

Поиск параметров модели по принципу максимума правдоподобия

$$p(X|W, M, S, \boldsymbol{\pi}) \to \max_{W,M,S,\boldsymbol{\pi}}$$

может быть осуществлен с помощью EM-алгоритма. E-шаг:

$$p(\boldsymbol{x}_n|\boldsymbol{z}_n) = \mathcal{N}(\boldsymbol{x}_n|W\boldsymbol{\mu}_{\boldsymbol{z}_n}, \Lambda + WV_{\boldsymbol{z}_n}W^T),$$
$$p(\boldsymbol{t}_n|\boldsymbol{z}_n, \boldsymbol{x}_n) = \mathcal{N}(\boldsymbol{t}_n|\Sigma_n(W^T\Lambda^{-1}\boldsymbol{x}_n + V^{-1}_{\boldsymbol{z}_n}\boldsymbol{\mu}_{\boldsymbol{z}_n}), \Sigma_n),\ \Sigma_n = (W^T\Lambda^{-1}W + V^{-1}_{\boldsymbol{z}_n})^{-1},$$
$$p(\boldsymbol{x}_n) = \sum_{\boldsymbol{z}_n} p(\boldsymbol{x}_n|\boldsymbol{z}_n)p(\boldsymbol{z}_n),$$
$$p(\boldsymbol{z}_n|\boldsymbol{x}_n) = \frac{p(\boldsymbol{x}_n|\boldsymbol{z}_n)p(\boldsymbol{z}_n)}{p(\boldsymbol{x}_n)},$$
$$p(\boldsymbol{t}_n|\boldsymbol{x}_n) = \sum_{\boldsymbol{z}_n} p(\boldsymbol{t}_n|\boldsymbol{z}_n, \boldsymbol{x}_n)p(\boldsymbol{z}_n|\boldsymbol{x}_n).$$

Здесь под суммой $\sum_{\boldsymbol{z}_n}$ понимается сумма по всем наборам компонент смесей $\sum^{K_1,\ldots,K_d}_{z_1,\ldots,z_d=1}$.

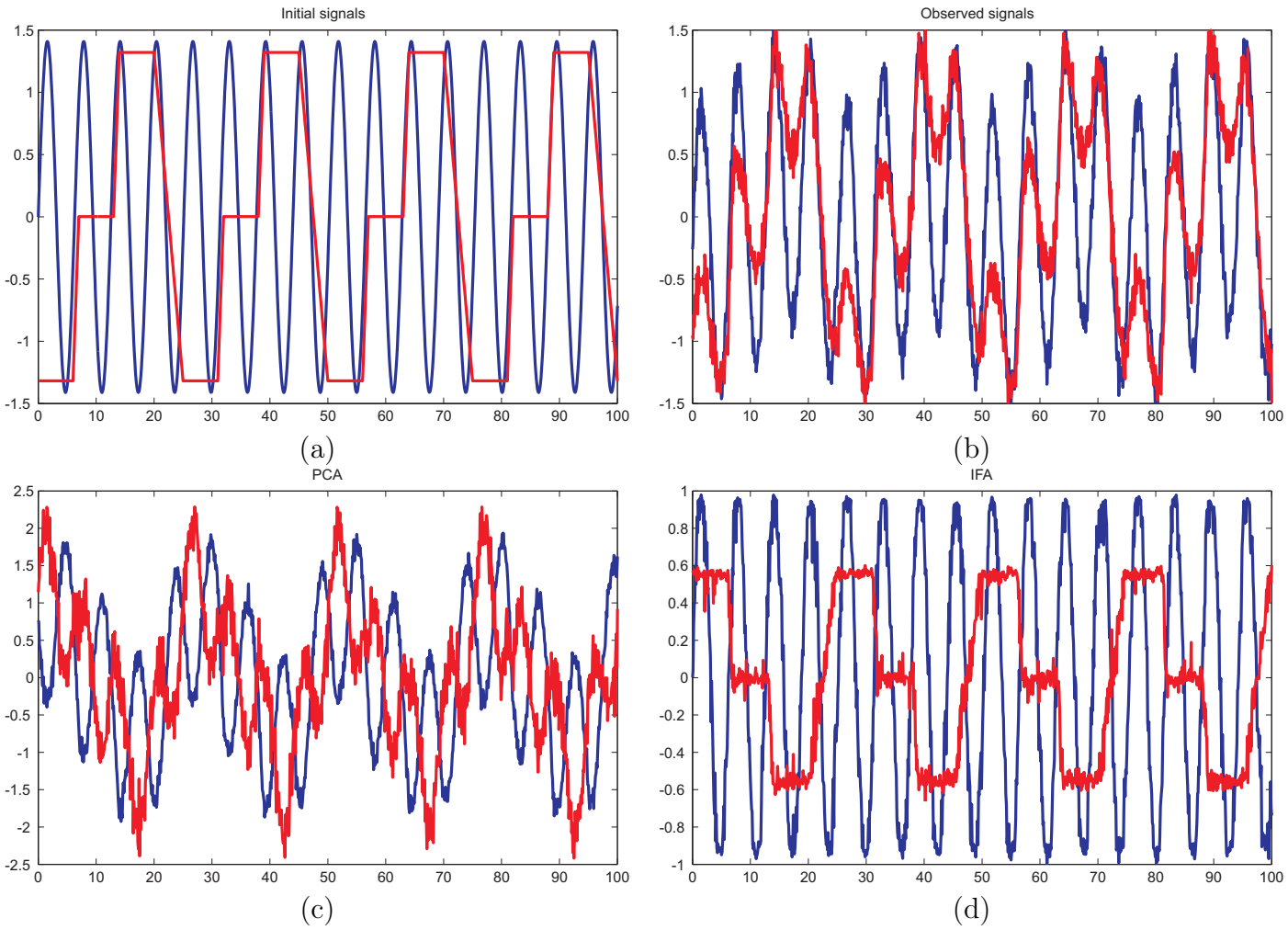


Рис. 12: Иллюстрация применения метода главных компонент и анализа независимых факторов для задачи разделения двух независимых источников. На рис. a показаны исходные сигналы, на рис. b – наблюдаемые сигналы, на рис. c показан результат применения метода главных компонент, а рис. d отражает результат применения анализа независимых факторов.

M-шаг:

$$W = \left(\sum_{n=1}^{N} \boldsymbol{x}_n \mathbb{E}_{\boldsymbol{t}_n|\boldsymbol{x}_n} \boldsymbol{t}_n^T\right) \left(\sum_{n=1}^{N} \mathbb{E}_{\boldsymbol{t}_n|\boldsymbol{x}_n} \boldsymbol{t}_n \boldsymbol{t}_n^T\right)^{-1},$$

$$\Lambda = \frac{1}{N} \sum_{n=1}^{N} [\boldsymbol{x}_n \boldsymbol{x}_n^T - 2\boldsymbol{x}_n \mathbb{E}_{\boldsymbol{t}_n|\boldsymbol{x}_n} \boldsymbol{t}_n^T W^T + W \mathbb{E}_{\boldsymbol{t}_n|\boldsymbol{x}_n} \boldsymbol{t}_n \boldsymbol{t}_n^T W^T],$$

$$\mu_j^i = \frac{\sum_{n=1}^{N} \sum_{\{z^k\}_{k\neq i}} \mathbb{E}_{\boldsymbol{t}_n|\boldsymbol{z}_n(i\leftarrow j),\boldsymbol{x}_n} t_{ni} p(\boldsymbol{z}_n(i \leftarrow j)|\boldsymbol{x}_n)}{\sum_{n=1}^{N} \sum_{\{z^k\}_{k\neq i}} p(\boldsymbol{z}_n(i \leftarrow j)|\boldsymbol{x}_n)},$$

$$s_j^i = \frac{\sum_{n=1}^{N} \sum_{\{z^k\}_{k\neq i}} p(\boldsymbol{z}_n(i \leftarrow j)|\boldsymbol{x}_n)[\mathbb{E}_{\boldsymbol{t}_n|\boldsymbol{z}_n(i\leftarrow j),\boldsymbol{x}_n} t_{ni}^2 - 2\mathbb{E}_{\boldsymbol{t}_n|\boldsymbol{z}_n(i\leftarrow j),\boldsymbol{x}_n} t_{ni}\mu_j^i + (\mu_j^i)^2]}{\sum_{n=1}^{N} \sum_{\{z^k\}_{k\neq i}} p(\boldsymbol{z}_n(i \leftarrow j)|\boldsymbol{x}_n)},$$

$$\pi_j^i = \frac{\sum_{n=1}^{N} \sum_{\{z^k\}_{k\neq i}} p(\boldsymbol{z}_n(i \leftarrow j)|\boldsymbol{x}_n)}{\sum_{l=1}^{K_i} \sum_{n=1}^{N} \sum_{\{z^k\}_{k\neq i}} p(\boldsymbol{z}_n(i \leftarrow l)|\boldsymbol{x}_n)}.$$

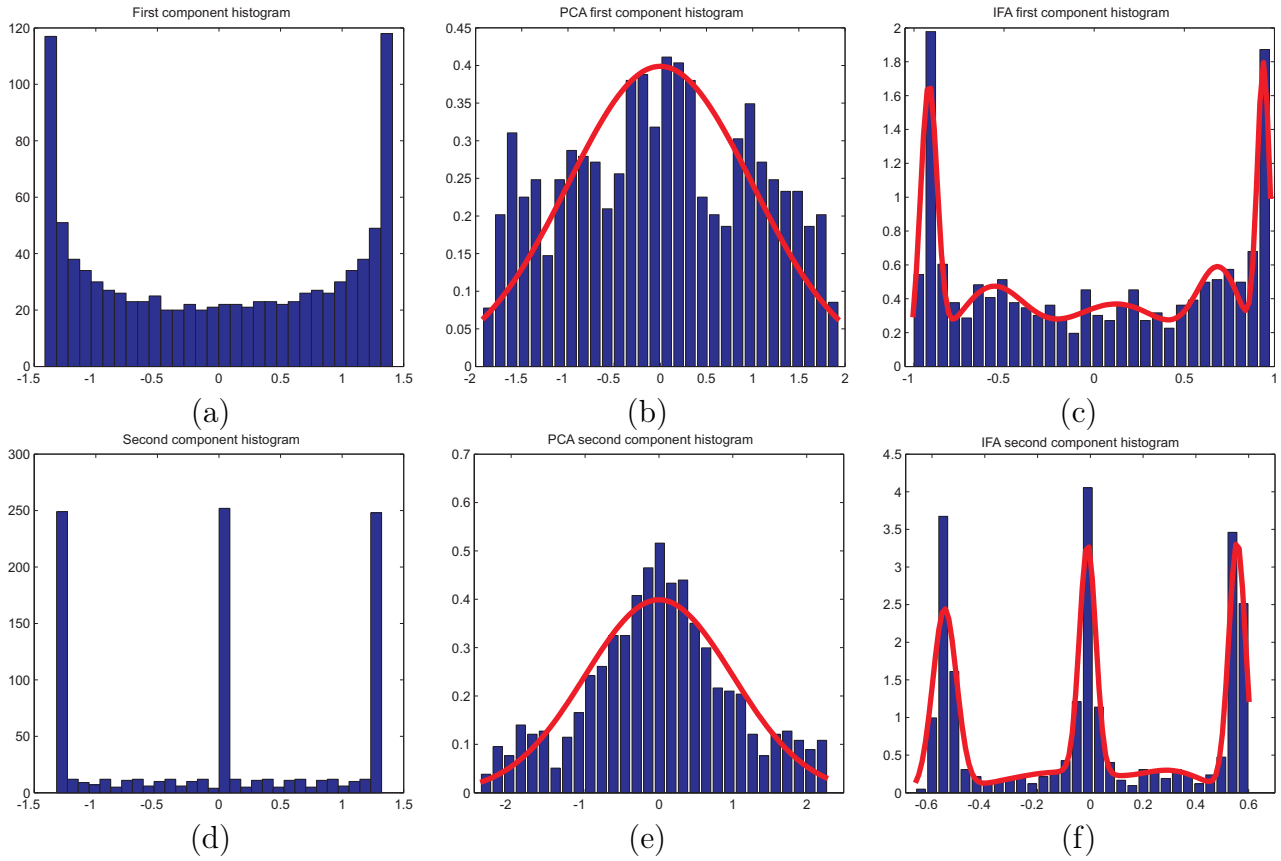


Рис. 13: Гистограммы для эксперимента, показанного на рис. 12. В первой строке представлены гистограммы распределения первой компоненты для истинного сигнала (a), метода главных компонент (b) и анализа независимых факторов (c). Во второй строке представлены аналогичные гистограммы для второй компоненты (d-f).

Здесь под символом $\boldsymbol{z}_n(i \leftarrow j)$ понимается такой вектор $\boldsymbol{z}_n$, что $z_{ni} = j$, а под суммой $\sum_{\{z^k\}_{k \neq i}}$ понимается сумма по всем компонентам z_j кроме i-ой, т.е. $\sum_{z_1=1}^{K_1} \cdots \sum_{z_{i-1}=1}^{K_{i-1}} \sum_{z_{i+1}=1}^{K_{i+1}} \cdots \sum_{z_d=1}^{K_d}$.

Рассмотрим модельный пример применения анализа независимых факторов. Пусть имеется задача разделения двух независимых источников, где исходные сигналы представлены на рис. 12a. Наблюдаемые сигналы образуются путем смешивания исходных сигналов с матрицей

$$W = \begin{bmatrix} 0.8 & 0.3 \\ 0.2 & 0.7 \end{bmatrix}$$

и добавления небольшого гауссовского шума (см. рис. 12b). Использование метода главных компонент приводит к результату, показанному на рис. 12c. Как и следовало ожидать, метод главных компонент не смог выделить исходные независимые сигналы. Применение анализа независимых факторов с параметрами $K_1 = K_2 = 5$ (см. рис. 12d), напротив, позволило найти компоненты сигнала, которые совпадают с истинными с точностью до масштаба и умножения на -1. Действительно, уменьшение дисперсии скрытой компоненты $\boldsymbol{t}$ всегда может быть скомпенсировано соответствующим увеличением матрицы смешивания W. Поэтому модель «анализ независимых факторов» не может определить дисперсию истинных скрытых компонент.

Причины различных результатов метода главных компонент и анализа независимых факторов для модельной задачи можно проиллюстрировать с помощью гистограмм (см. рис. 13). Видно, что гистограммы истинных сигналов далеки от стандартного нормального. Поэтому метод главных компонент находит матрицу смешивания, отличную от истинной, чтобы обеспечить нормальность распределения своих скрытых компонент (как уже отмечалось выше, это всегда можно сделать). Анализ независимых факторов благодаря более гибкой модели $p(\boldsymbol{t})$ успешно справился с задачей и, в частности, обнаружил двумодальность и трехмодальность для первой и второй скрытой компоненты соответственно.

Список литературы

[1] J.J. Sylvester. On the reduction of a bilinear quantic of the nth order to the form of a sum of n products by a double orthogonal substitution // Messenger of Mathematics, 19, 1889, pp. 42–46.

[2] K. Pearson. On lines and planes of closest fit to systems of points in space // Philosophical Magazine 2, 1901, pp. 559–572.

[3] M.E. Tipping, C.M. Bishop. Probabilistic Principal Component Analysis // Journal of the Royal Statistical Society, B, 61(3), 1999, pp. 611–622.

[4] C.M. Bishop. Pattern Recognition and Machine Learning. Springer, 2006.

[5] Д.П. Ветров, Д.А. Кропотов, А.А. Осокин. Автоматическое определение количества компонент в EM-алгоритме восстановления смеси нормальных распределений // Ж. вычисл. матем. и матем. физ., 2010, т. 50, № 4, с. 1–14.

[6] A. Hyvärinen, J. Karhunen, E. Oja. Independent Component Analysis. Wiley, 2001.

[7] M.E. Tipping, C.M. Bishop. Mixtures of Probabilistic Principal Component Analysers // Neural Computation 11(2), 1999, pp. 443–482.

[8] H. Attias. Independent Factor Analysis // Neural Computation 11(4), 1999, pp. 803–851.